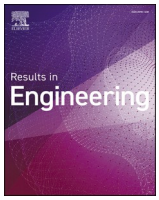
Results in Engineering 22 (2024) 102132

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering



Review article

A systematic review of big data innovations in smart grids

Hamed Taherdoost [a,b,*]

[a] University Canada West, Vancouver, Canada
[b] GUS Institute, Global University Systems, London, UK

ARTICLE INFO

Keywords:
Data science
Smart environment
Big data analytics
Energy management
Demand response

ABSTRACT

Multiple industries have been revolutionized by the incorporation of data science advancements into intelligent environment technologies, specifically in the context of smart grids. Smart grids offer a dynamic and efficient framework for the management and optimization of electricity generation, distribution, and consumption, thanks to developments in big data analytics. This review delves into the integration of Smart Grid applications and Big Data analytics by reviewing 25 papers screened with PRISMA standard. The paper matter encompasses critical domains including adaptive energy management, canonical correlation analysis, and novel methodologies including blockchain and machine learning. The paper emphasizes contributions to energy efficiency, security, and sustainability by means of a rigorous methodology.

1. Introduction

Information and communication technologies (ICT) are ubiquitous in virtually every aspect of contemporary society [1]. As the reliance of the healthcare system on technology increases, health science students are required to enhance their ICT proficiency [2]. ICT facilitates the implementation of innovative learning materials and approaches, allowing for increased student collaboration and the concurrent acquisition of technological expertise [3]. Notwithstanding the absence of a universally accepted definition, ICT is generally understood to encompass all hardware, software, systems, and applications that enable organizations and individuals to communicate in the digital realm [4]. The consequential increase in data volumes, both within governmental organizations and private enterprises, has presented researchers with multifaceted challenges. Key among these challenges is the need to develop effective solutions for the manipulation and analysis of large datasets, as well as the establishment of mechanisms to seamlessly transmit these data from one site to another.

The interconnection of these technologies generates an automated ecosystem in which data is collected by the Internet of Things (IoT) devices and subsequently processed and analyzed through the utilization of big data analytics and artificial intelligence algorithms [5,6]. By transforming enormous data sets from diverse origins, such as the IoT, into a coherent structure, big data analytics enables businesses to acquire valuable insights and formulate decisions based on data [7,8]. The IoT, which is comprised of sensors and interconnected devices, is a crucial enabler of intelligent environments by enabling the collection of real-time data that can be analyzed to detect trends and patterns [9,10]. In, particular, smart cities benefit from the utilization of big data analytics to detect resource waste, optimize resource consumption, and enhance energy management. As a result, residents enjoy greater efficiency, sustainability, and a higher standard of living [11–13]. The application of ICT in the energy sector transformed the current grid in the era of Industry 4.0 [14].

Several studies examine advancements in data science as they pertain to intelligent environment technologies. To automate city systems, Sarker [15] emphasized the significance of extracting valuable knowledge from city-data. Ullah et al. [16] examined the significance of data in driving innovation and development and the role of machine learning and the IoT in realizing a data-centric smart environment. The article by Grossi et al. [17] discussed the capacity of data science to bring about disruptive innovation in diverse sectors, such as energy and environment. It did so by facilitating the collection of high-resolution data and augmenting the beneficial effects on the environment. Atitallah et al. [18] described how data acquisition and processing utilizing a variety of technologies, including IoT and deep learning, can benefit a smart city. As a whole, these papers offer valuable perspectives on the application of data science and associated technologies to foster advancements in intelligent environment technologies, with a specific focus on smart cities and environmental sustainability.

The examination of practical applications of data science in the establishment of environmentally sustainable smart cities, the

* Corresponding author. University Canada West, Vancouver, Canada.
E-mail address: hamed.taherdoost@gmail.com.

https://doi.org/10.1016/j.rineng.2024.102132
Received 30 December 2023; Received in revised form 5 April 2024; Accepted 10 April 2024
Available online 21 April 2024

integration of big data, artificial intelligence, and IoT technologies to tackle climate change and sustainability issues, the creation of data-driven models for automated city systems, and the investigation of data science's potential to support unified approaches are all areas of research that require further investigation when undertaking a review of data science innovations in smart environment technologies [15,17,19–21]. The aforementioned gaps underscore the criticality of connecting progress in data science theory to practical implementations in smart environment technologies to tackle urgent environmental and societal issues.

This article examines the most recent advancements in the application of big data to smart grids. It concludes with a systematic literature review of big data analytics and its critical role in determining the trajectory of smart grid technologies in the future. A comprehensive analysis of established research methodologies, significant discoveries, and discussions comprise the systematic review, which provides an all-encompassing glimpse into the present state of knowledge in this field.

2. Literature background

2.1. Data science in modern societies

The influence of data science on contemporary societies is substantial, as supported by an assortment of scholarly articles. Data science is a composite field of study that has the potential to enhance scientific research, government administration, business decision-making, and innovation in industry, science, and policy [17,22]. The importance of data science in addressing societal issues such as climate change, urban planning, and healthcare is emphasized, as it enables the creation of data-driven solutions to complex problems [15,21,23]. Data science is depicted as a transformative force capable of introducing disruptive innovations across multiple spheres of society [17].

Data science is an all-encompassing and interdisciplinary paradigm that combines various models and theories to convert data into knowledge (and value). In addition to validating established theories and models, experiments and analyses conducted on enormous datasets facilitate the identification of patterns that emerge from the data. Such insights can assist scientists in developing more accurate theories and models, ultimately leading to a more comprehensive comprehension of the intricate nature of social, economic, biological, technological, cultural, and natural phenomena. When accessible data is reinterpreted for analysis, the outcomes of data science are not in line with the initial motivations that drove data collection. The aforementioned elements are collectively influencing a transformation in the scientific method, research, and societal decision-making processes [24].

Data science is founded upon three confirming facts: the emergence of big data, which offers a substantial quantity of real-world examples from which to learn; the progressions in data analysis and learning methodologies that enable the derivation of predictive models and behavioral patterns from big data; and the developments in high-performance computing infrastructures that facilitate the ingestion, management, and execution of intricate analyses on big data [25]. Through the use of deep learning and reinforcement learning techniques, artificial intelligence offers practical solutions for optimal decision making on a global scale [26].

Data science is a dynamic and multidisciplinary discipline that emerges from the convergence of mathematics, statistics, information theory, computer science, and social science (Fig. 1). By integrating these disciplines, a solid and mutually beneficial framework is established to tackle the difficulties that arise from handling extensive and intricate datasets. Data scientists develop exhaustive methodologies, models, and algorithms by capitalizing on the respective merits of computer science, mathematics, statistics, information theory, and social science. The interdisciplinary character of data science not only guarantees technical expertise but also takes into account the wider societal framework, rendering it a potent instrument for furthering knowledge and catalyzing constructive transformations in various fields.

2.2. Domains within data science

Data Science is an interdisciplinary domain that incorporates methodologies and concepts from numerous other fields, thereby enhancing its holistic approach to data management and insight extraction [27,28]. Domain knowledge and statistics, computer science, and programming are instrumental in the development and implementation of techniques for efficient data analysis. Data scientists analyze data sets utilizing an array of statistical and analytic methods, such as classification, regression, anomaly detection, and others [29]. Data analysts have access to a variety of data analysis techniques, such as sentiment analysis, cohort analysis, cluster analysis, and time series analysis [30–33], as the discipline of data science expands at an accelerated rate.

The systematic extraction of knowledge from data, also known as data science, has garnered considerable interest in recent times [34]. As one might expect, data science is at the forefront of a paradigm shift in science [35]. In numerous fields of study, its epistemological assumptions, challenges, and opportunities have been examined [36,37]. Concerns remain, however, as to whether this represents the resurgence of empiricism a genuine fourth paradigm of science [38], or merely an expansion of established paradigms accompanied by novel instruments and approaches to scientific investigation [39].

Approaching is the fourth industrial revolution. At present, a considerable number of enterprises are adopting Industry 4.0, capitalizing on nascent developments in automation, the industrial IoT, big data, and cloud computing [40]. The emergence of the industrial IoT has facilitated the real-time collection of an unprecedented volume of data by sensors integrated into networked physical devices. This data

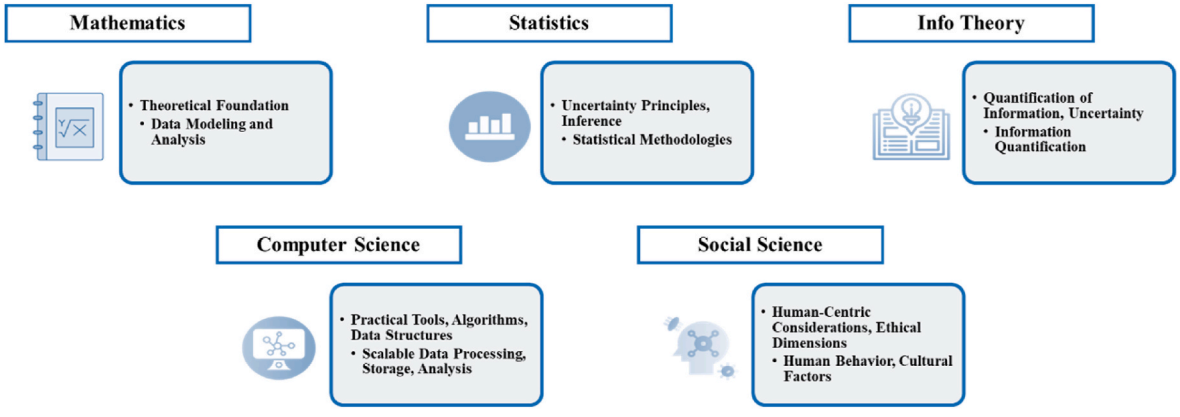


Fig. 1. Disciplines in data science.

empowers manufacturing operations, processes, and systems to attain notable improvements in productivity, efficiency, and self-management [41,42]. Proficiency in machine learning, artificial intelligence, and data analytics techniques is essential in the manufacturing sector [43–45]. ρReveal, a novel Big Data Analytics (BDA) scheme for predicting energy prices in Smart Grids, was introduced by Kumari and Tanwar [46]. It makes use of Spark-based analytics for load reduction techniques and an artificial intelligence-based Bidirectional Long Short-Term Memory (BiLSTM) model for precise price forecasting. In terms of data security and prediction accuracy (RMSE, MAE, and MAPE), Reveal performed better than previous methods.

2.3. Smart grid management and operation

Demand-side management techniques, such as demand-response control in smart grids, specify how the demand side responds to price strategies or incentive actions from a power unit [47,48]. Supply-demand balance can be promoted and cost-effective, high--quality, customized services can be provided to demand-side consumers with the help of demand-response management [49].

By analyzing historical and real-time consumption patterns, utilities can improve their demand response strategies to meet consumer and grid needs. Better load forecasting, demand-side management, and consumer participation can result. The emergency demand response (EDR) pilot in southwestern China during a heatwave was successful. The study found that incentive-based EDR policies like time-of-use (TOU) pricing can reduce peak loads and demand significantly. The study also found that smart thermostats and home automation systems improve price-based demand responses [50].

To enable proactive maintenance and minimize downtime, predictive maintenance employs data and analytics to forecast when a component in an actual system is likely to fail [51]. In the context of Industry 4.0, where technical system complexity is constantly increasing, this approach is especially helpful [52]. It is especially helpful in anticipating unplanned deterioration, which enables operators to take preventative measures and stop malfunctions before they happen [53].

The large amount of data generated by smart grids will help utilities understand customer conservation, consumption, and demand, track downtime, and power failures. This will be difficult for utilities without the systems and data analysis skills to handle these data. Therefore, utilities now aim to manage high-volume data and use advanced analytics to turn data into knowledge and actionable plans [54].

The capacity of big data analytics to absorb, process, and analyze massive amounts of data in almost real-time is one of its main advantages for smart grids. Real-time data collection is made possible by Supervisory Control and Data Acquisition (SCADA) technologies, which also allow for remote grid performance monitoring, control, and analysis to increase operational dependability and efficiency. Automation of the grid depends on SCADA and Distribution Management Systems (DMS), which integrate sophisticated algorithms and analytics to manage grid assets, identify faults, and optimize energy flow for stable and balanced power distribution. Energy Management Systems (EMS) are crucial to grid automation because they facilitate load management and grid balancing, regulate and optimize energy resources, and monitor energy generation, consumption, and storage. With EMS technologies, grid operators can optimize energy use, integrate renewable energy sources, and improve stability for a dependable power supply [55,56].

Distributed Energy Resources, a product of technological advancement, has raised consumer involvement in energy generation and management within the Smart Grid system [57]. There are additional difficulties in integrating big data analytics with smart grids, such as privacy and cyber security concerns. Ensuring the security and privacy of this data is becoming more and more crucial as the volume of data produced by smart grids increases [58]. End users can benefit from a multitude of services provided by a smart grid system, including energy trading (ET), load forecasting, and load management. Since data in an SG environment moves between various devices via an open channel—the Internet—security and privacy are perpetually difficult problems. Despite the fact that there are numerous solutions to this issue in the literature, they are insufficient to address security, privacy, latency, or real-time ET settlement [59]. For peer-to-peer trading in smart grid systems, Kumari et al. [59] proposed ET-DeaL, a Secure Energy Trading scheme based on Smart Contracts. ET-DeaL surpasses conventional systems in performance metrics by utilizing IPFS and Ethereum smart contracts to manage energy loads efficiently in the residential, industrial, and electric vehicle sectors while addressing security, privacy, and latency concerns.

3. Theoretical concepts

The utilization of High-performance computing has become prevalent across numerous scientific and practical domains. High-performance computing techniques comprise technologies and methods used to accelerate the execution of complex computational processes. Hardware accelerators, parallel processing, cloud computing, distributed computing, multi-core processing, and cluster computing are all prevalent high-performance computing techniques. These types of systems frequently comprise numerous computing nodes and are frequently linked via a high-bandwidth network. By effectively managing vast quantities of data, decomposing complex problems into smaller components, and concurrently resolving each component, high-performance computing systems achieve overall processing times that are considerably quicker. Extremely complex problems in the domains of aerospace [60], Fintech [61], chemistry [62], biology [63], physics [64], and others could be resolved with the assistance of high-performance computing techniques. The integration of modeling, algorithm development, software construction, and computational simulation through the application of high-performance computing has emerged as an indispensable foundation in the realm of cutting-edge fundamental science research.

The use of high performance computing resources has been shown to lead to shorter runtimes across all categories, demonstrating its consistent trend in enhancing computational efficiency [65]. Furthermore, high performance computing is described as having computing power that meets the needs of an intelligent society and serves as infrastructure like water and electricity [66]. These findings underscore the significance of advanced algorithms customized to the unique attributes of smaller, intricately structured datasets, which enable instantaneous processing and enhance the nuanced nature of problem-solving.

The capability of high-performance computing to process lesser datasets in real time is a critical attribute for applications including scientific simulations, engineering scenarios, and financial modeling. The algorithms have been meticulously optimized to capitalize on the intrinsic structure of the data, delivering results and insights virtually in real time. Due to its remarkable flexibility, high-performance computing is an essential resource in both scientific research and industrial applications that require precise curation of datasets. High-performance computing is a preferred instrument in material science [67], drug discovery [68], and computational fluid dynamics [69], demonstrating its adaptability and effectiveness when confronted with smaller, highly structured datasets. This can be observed in various studies and applications, such as designing high-performance computing and big data converged systems, also referred to as High-Performance Data Analytics (HPDA), which is a hard task requiring careful placement of data [70].

One notable attribute of high-performance computing is its heavy reliance on substantial computation. Standard computing platforms may find it impracticable or overly time-consuming to process large datasets, implement complex mathematical models, and conduct simulations. For these tasks, high-performance computing architectures are meticulously designed. The prioritization of substantial computation holds particular significance in domains including financial modeling, molecular

dynamics simulations, weather prediction, and other applications that demand substantial computational resources. Efficient and precise computations are necessary in these domains to extract significant insights, facilitate well-informed decision-making, and simulate real-world phenomena precisely [71].

Within the domain of data science and the wider realm of technological advancement, the interdependence of data processing and high-performance computation becomes conspicuous. High-performance computing enables data scientists and researchers to extract valuable insights from extensive datasets, thereby promoting progress in diverse domains such as engineering simulations, scientific inquiry, and machine learning. The necessity for computational capacity to manipulate, analyze, and derive meaningful patterns from these enormous datasets becomes ever more critical as their size and complexity continue to increase [72].

3.1. Emergence of data science

The advent of Data Science has brought about substantial changes in the way we confront the difficulties presented by the growing dependence on ICT. Conventional High-Performance Computing techniques, which were previously indispensable for smaller, structured datasets, encountered constraints when confronted with prodigious, unstructured data. The article by Barakat et al. [73] highlighted the significance of Data Science in tackling the difficulties associated with conditions such as ARDS. This article examined a range of Data Science methodologies, such as mechanistic modeling, deep learning, and time series analysis, with a particular emphasis on the integration of High-Performance Computing to facilitate efficient algorithmic support.

The paper by Mellone et al. [74] presented a paradigm shift in environmental research by addressing the changing demands of the field through the implementation of High-Performance Cloud-Native Computing. This methodology took advantage of scalable high-performance computation in the cloud, showcasing its advantages in terms of resource conservation and enhanced performance. An additional article by Oujja et al. [75] emphasized the critical nature of ICT instruments utilized to analyze genomic data, with a specific emphasis on SARS-CoV-2. It introduced a data science-driven high--performance computing-based instrument, emphasizing the importance of computational capability in the context of RNA clustering for virus mutation analysis.

The integration of data science and high-performance computing is apparent in numerous fields, including healthcare, as elaborated in the article by Courneya and Mayo [76]. By bridging the distance between computational infrastructure and wet lab setups, this article offers bioinformatics and high-performance computing support to researchers. Moreover, Belov et al. [77] demonstrated this convergence also benefits the educational sector. They examined the characteristics and attributes of high-performance computing platforms that are employed in educational procedures, with a particular focus on the significance of having access to suitable tools and resources for data science and parallel programming instruction.

A fundamental attribute of Data Science is its emphasis on extracting insights and knowledge from enormous datasets, thereby revealing patterns, trends, and valuable information that may remain concealed when utilizing conventional analytical methods. This paradigmatic change has not only resulted in the redefinition of data management methodologies, but has also fostered the development of innovative theories and techniques that contribute to the progress of numerous scientific, industrial, and governmental sectors. The papers that discuss the domain-spanning collaborative impact of high-performance computing and data science are listed in Table 1.

The management of unstructured datasets is a critical aspect of Data Science, presenting both a substantial challenge and an opportunity in the current data landscape. Unstructured data encompasses diverse formats, including sensor-generated information, text, images, and videos. To navigate unstructured datasets and extract meaningful information, Data Science employs a range of sophisticated techniques such as machine learning algorithms, natural language processing, computer vision, and deep learning. These advanced methods enable Data Science to uncover valuable insights from unstructured data, turning what might seem like disorder into a source of hidden patterns and knowledge [31,82,83].

Furthermore, the ability of Data Science frameworks to scale facilitates the effective manipulation of enormous datasets. Through the utilization of Big Data technologies, distributed computing, and parallel processing, Data Scientists are empowered to manage and analyze enormous volumes of data that surpass the capabilities of conventional computing systems. Scalability is an essential factor in effectively managing the complexities presented by the exponential increase in data produced by contemporary technologies such as the IoT and interconnected smart environments. Table 2 presents an extensive assortment of methodologies and applications, thereby offering valuable insights into the intricate convergence of data science, big data, and intelligent environments.

3.2. Processes and approaches for knowledge extraction

Knowledge extraction is the process of extracting relevant and meaningful information from unstructured or structured data sources, such as text, documents, images, and relational databases [93,94]. The process includes identifying patterns and relationships, transforming them into actionable knowledge [95]. Techniques like named entity recognition [96], text mining [97], natural language processing [97], information retrieval [98], and machine learning [99], including deep learning [100], rule-based systems [101], decision trees [102], and neural networks [103], contribute to this extraction.

Knowledge extraction has various applications, including sentiment analysis, text summarization, and information retrieval [104]. The resulting knowledge needs to be in a machine-readable and machine-interpretable format and needs to be able to be queried. Fig. 2 depicts a closed-loop knowledge extraction process with implementation tips, covering data collection to actionable insights.

Table 1
Impact of data science on high-performance computing.

Study	Collaboration of two concepts	Healthcare Applications	Environmental Modeling	Bioinformatics and -omics	Genomic Data Analysis	Educational Processes	Quantum Computing Integration
[78]	✓						
[73]	✓	✓		✓			
[74]	✓		✓				
[76]	✓	✓		✓			
[75]	✓				✓		
[79]	✓	✓					
[80]	✓						
[77]	✓					✓	
[81]	✓						✓

Table 2
Exploration of data science and big data applications in smart environments.

Study	Technologies/ Frameworks	Application	Evaluation
[84]	IoT, Intrusion Detection, Security Protocol	Real-time security system, Communication protocol	Security analyses, Efficiency evaluation
[85]	IoT, AllJoyn, MongoDB, Storm	Monitoring, Management, Big Data analytics	Experimental results, Performance evaluation
[86]	Blockchain, Decentralized Auditing	Ensuring integrity and auditability of big data	Theoretical analysis, Experimental evaluation
[87]	VRGIS, TIN Data Model	Intelligent tourism service system, VRGIS applications	Experiment results, User experience
[88]	Spark, Hadoop	Real-time processing, Smart transportation planning	Throughput, Validation with authentic data
[89]	Kalman Filter, Weighted Hybrid Recommender System	Enhanced living environment, User behavior prediction	Recall and precision rates
[90]	MapReduce, Partial Order Reduction	Attribute reduction for power systems, Simulation examples	Performance observed through a Hadoop platform
[91]	Fuzzy Logic, Multi-Fuzzy Agent-based WSN	Noise reduction, Smart data extraction	Simulation results
[92]	Blockchain, PKI/CA Security System	Decentralized trust service system, Smart City development	Evaluation model, Vertical comparison

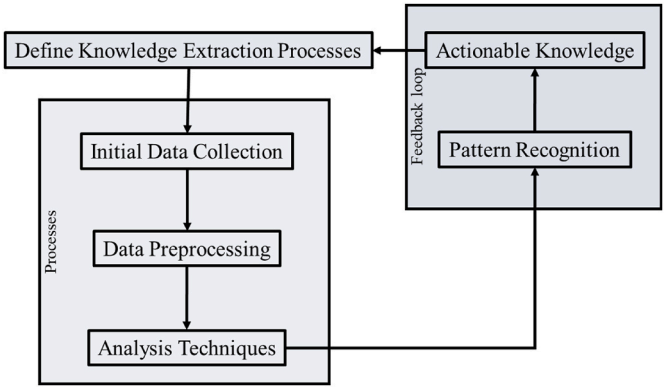


Fig. 2. Iterative knowledge extraction loop: implementation tips.

4. Methodology for systematic review

The search strategy involves employing a comprehensive set of keywords and search terms, including terms related to Big Data, Smart Grids, and associated technologies, such as "Big Data" AND "Smart Grid". The search was conducted across Scopus and Web of Science database. This strategy aims to capture studies focusing on the integration of Big Data in Smart Grid technologies.

- Research Questions (RQs) for the Review:

RQ1How do different big data analytics approaches contribute to the optimization and efficiency of smart grid technologies?

RQ2What are the challenges and solutions related to data security and privacy in the integration of big data analytics in smart grids?

RQ3How do machine learning and artificial intelligence techniques enhance decision-making processes in smart grid applications?

RQ4What role does blockchain technology play in ensuring secure and transparent transactions within the context of energy smart grids?

4.1. Inclusion and exclusion criteria

The criteria for selecting relevant studies include a publication timeframe between 2019 and 2023 to ensure currency, with a specific focus on the impact of Big Data in Smart Grid technologies. Only English articles will be considered for inclusion. Exclusion criteria encompass reviews, book chapters, conference papers, and books, aiming to prioritize primary research articles that directly contribute to the understanding of the integration, challenges, and outcomes of Big Data in the context of Smart Grids.

4.2. Data extraction

The data extraction process will be systematic and transparent, employing a standardized data extraction form. Key variables to be extracted include study design, methodologies utilized, and key findings. Additionally, variables relevant to Smart Grid technologies and their interaction with Big Data will be identified. The extraction form will be designed to capture essential information from each included study, ensuring consistency in the retrieval of data.

4.3. Data synthesis

The synthesis of data involved a systematic collation of findings from the included studies. A structured approach will be employed to present key themes and trends emerging from the literature. The data synthesis will encompass a narrative summary of study designs, methodologies employed, and key outcomes, with a specific focus on the interaction between Big Data and Smart Grid technologies.

4.4. Data selection

Following the initial search on December 28, 2023, 444 articles were identified. After applying inclusion and exclusion criteria, 32 papers were retained. Following a detailed review of abstracts and scopes, 24 papers remained. An additional paper was found through further search, resulting in a final selection of 25 papers for the systematic review (Fig. 3).

5. Results

A total of 24 studies related to various aspects of smart grids, big data analytics, and related technologies were identified. The studies cover a range of topics, including canonical correlation analysis, adaptive energy management, post-evaluation systems, big data collection and utilization in smart factories, integration of big data and blockchain, demand response management, electricity theft detection, energy consumption prediction, big data compression, outlier rejection for load forecasting, fog computing, differentially private clustering, non-technical loss fraud detection, evaluation of big data frameworks, anonymous batch verification, optimal big data processing, temporal-functional-spatial big data computing, and robust big data analytics for electricity price forecasting.

The keyword cloud from our selected articles as shown in Fig. 4, indicates a strong focus on integrating Big Data analytics into Smart Grid technologies. Terms like "Big Data," "Smart Grid," and "Smart Power Grids" underscore the core areas of research. Additionally, there is a growing interest in leveraging "Data Analytics," "Data Mining," and "Internet of Things (IoT)" for extracting insights from Smart Grid data. The keywords "Fog Computing," "Green Computing," and "Energy Efficiency" suggest a trend toward exploring innovative computing paradigms and sustainable practices. The recurring focus on "Electricity Theft Detection," "Data Privacy," and "Security and Privacy" highlights the rising importance of securing Smart Grids and ensuring data confidentiality.

The distribution of subject areas among the 25 papers reflects a

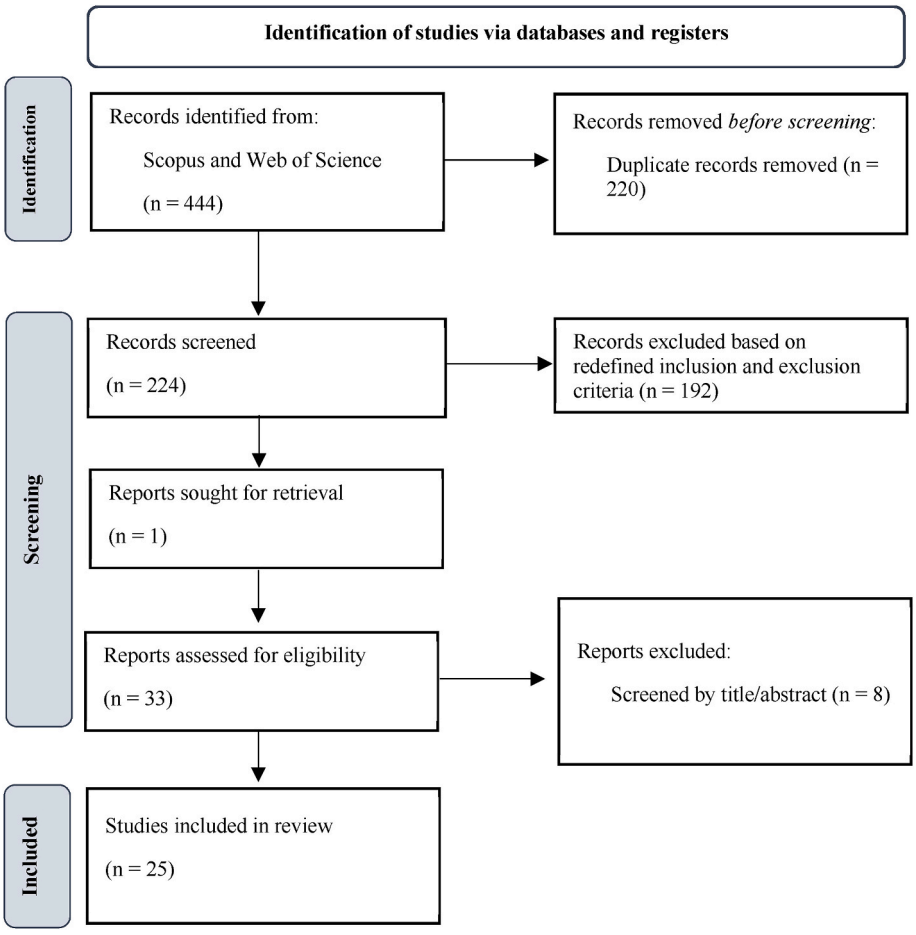


Fig. 3. PRISMA literature search methodology flowchart.

Fig. 4. Cloud of selected article keywords (created by www.wordclouds.com.

predominant focus on Computer Science, comprising 76 % of the research output (Fig. 5). This emphasis underscores the significance of computational methods, algorithms, and data analytics in the exploration of smart grids and related technologies. Engineering follows closely as the second most represented subject area, constituting 48 % of the papers, indicating a strong applied orientation with a focus on the

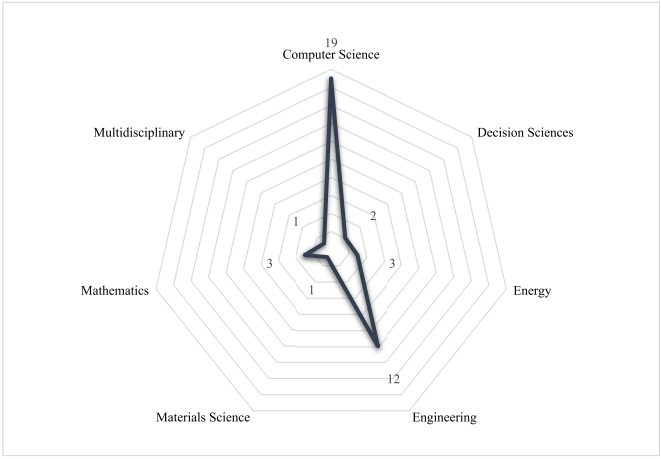


Fig. 5. Subject area distribution of selected papers.

development and implementation of technologies within the smart grid domain. The inclusion of subject areas such as Decision Sciences, Energy, Mathematics, and Materials Science reveals a multidisciplinary approach, acknowledging the diverse challenges inherent in the study of smart grids. The presence of papers in Energy suggests a concentrated consideration of energy-related aspects, while the representation of Decision Sciences and Mathematics implies a quantitative and analytical approach to problem-solving. Although less prevalent, the inclusion of Materials Science underscores the importance of materials in the context of smart grids, possibly pertaining to advancements in energy storage or grid infrastructure.

Table 3 serves as a quick reference for understanding the methodologies, datasets, and outcomes of diverse research efforts in the field of Smart Grids and Big Data analytics.

6. 6. discussion

RQ1. How do different big data analytics approaches contribute to the optimization and efficiency of smart grid technologies?

The effectiveness and productivity of smart grid technologies are notably impacted by a multitude of big data analytics methodologies, as supported by an extensive research body. Canonical correlation analysis is employed by scholars including Jiang et al. [105] to investigate the interrelationships between gas consumption, electricity consumption, and climate change. This method yielded significant insights into patterns of consumption. Comparing data processing methods, Gupta and Chaturvedi [106] examined Adaptive Energy Management in Smart Grids, placing particular emphasis on the dependable linear regression method, which achieves an impressive accuracy rate of 98 %. The integration of these methodologies aids in the comprehension of consumer behaviors and the resolution of energy management challenges.

By virtue of integrating big data analysis and microservice frameworks, Wang et al. [128] presented a post-evaluation platform for intelligent grid electricity generation in wind farms. This platform exemplifies the pragmatic implementation of big data in enhancing the efficacy of energy generation by assessing the operational status of wind farms and providing technical assistance. Furthermore, the TOTEM framework, which Jose et al. [109] put forth, merges blockchain technology with big data to facilitate the management of energy smart grids. This framework enhances data security and privacy while providing a reliable platform for energy transactions. The aforementioned studies underscore the significance of big data analytics in the context of secure energy management and post-evaluation systems.

Akram et al. [111] directed their attention towards novel methodologies in the realm of electricity theft detection. They demonstrated the efficacy of enhanced RUSBoost classifiers, which achieve accuracy rates of 91.5 % and 93.5 %, respectively. In the interim, Kumari et al. [110] presented a demand response management strategy that makes use of the Prophet model to illustrate the efficacy of energy consumption forecasting in smart grids. The aforementioned studies underscore the significance of big data analytics in augmenting security protocols and precisely forecasting energy requirements. An optimal singular value decomposition (SVD)-based large data compression approach for smart grids is proposed by Hashemipour et al. [113], which showcases effective compression levels. Intelligent optimization determines optimal values, improving data quality and compression ratio. Comparative analysis shows superior compression levels to existing SVD rank reduction methods, emphasizing the need for application-specific optimization for reliable performance.

The studies offer quantifiable metrics that can be utilized to assess the efficacy of big data analytics methodologies. Quantitative measures such as accuracy percentages, area under the curve (AUC), F1-score, precision, and recall are employed to evaluate the dependability and efficacy of these methodologies. For example, Gupta and Chaturvedi [106] provide accuracy percentages for different data processing methods, whereas Javaid et al. [114] evaluate the performance of their deep siamese network using AUC, F1-Score, precision, and recall. They presented an adaptive synthesis approach for electricity theft detection utilizing an imbalanced big data set and a deep siamese network. The utilization of these quantitative metrics provides a strong basis for analyzing the effects of big data analytics on the optimization of smart grid technologies, with a particular focus on the accuracy and dependability attained across diverse implementations. The accuracy rates of various big data analytics models for a variety of smart grid applications are compared in Table 4.

Boosting methods such as RUSBoost BSA and RUSBoost MRFO exhibit superior performance compared to SVM (71 %) and LR (63 %). CNN achieves an accuracy of 85.1 %, while boosted models can attain as high as 93 %. When combined with a CNN-LSTM deep siamese network, the ADASYN method achieves an accuracy of 95.3 %–83.9 % across a variety of training ratios. When ELF strategies and HORM are combined for outlier rejection, accuracy increases significantly from 87 % to 95 %. Moreover, feature selection techniques such as FBFS and FSSO4 contribute to improvements in precision, with the latter attaining 93 %. Linear regression demonstrates the highest degree of accuracy (98 %), surpassing logistic regression (96 %) and K-NN (92 %), among conventional methods. The GA-LSTM approach, which is founded on the Genetic Algorithm, attains a moderate level of accuracy, specifically 80.27 %–82.42 %, which is notably higher than that of random methods. In addition, the accuracy of hybrid feature selectors utilizing DE-based SVM classifiers surpasses 90 %. Although not directly associated with accuracy, innovative frameworks present encouraging improvements in efficiency, including a 95 % convergence ratio and 81 % bandwidth enhancement.

The progress made in accuracy has a direct influence on the efficiency of smart grid operations, guaranteeing enhanced dependability in the areas of load forecasting, energy management, and electricity theft detection. Proving the capability of robust big data analytics to optimize smart grid performance, the incorporation of sophisticated techniques such as hybrid feature selectors and Genetic Algorithm-based models further improves precision. Innovative computing frameworks specifically engineered for expansive smart grids not only enhance computational efficacy but also facilitate bandwidth conservation, thereby emphasizing the far-reaching consequences of precision advancements on the overall performance and dependability of the system.

RQ2. What are the challenges and solutions related to data security and privacy in the integration of big data analytics in smart grids?

The incorporation of big data analytics into smart infrastructure presents a range of complex obstacles, with data security and privacy being of particular significance. Significantly, the research conducted by

Table 3
Methods, datasets, and metrics of big data insights in smart grids.

Study	Methods	Dataset Size	Analysis Techniques	Key Findings	Performance Metrics
[105]	Canonical Correlation Analysis, Consumer Segmentation, Preprocessing	3 datasets (24 values/day, 1-year period)	Canonical correlation, Consumer segmentation, Comparison analysis	Overview, Typical patterns, Comparison on climate zones and locations	Not specified
[106]	Linear Regression, Logistic Regression, K-Nearest Neighbors	50,000 instances (smart meters), 10,000 attributes	Comparison of linear regression, logistic regression, K-Nearest Neighbors	Linear regression: 98 %, Logistic regression: 96 %, K-Nearest Neighbors: 92 %	Accuracy percentage
[107]	Microservice Framework, Big Data Analysis	Not specified	Evaluation of wind farms, Technical support, Visualization	Tested and proven for processing and analyzing massive data	Not specified
[108]	Apache Kafka, Big Data Ecosystem	Data from geographically remote, independent networks	Data gathering, Data convergence, Data analysis	Stable and effective exchange and collection of data using Apache Kafka	Not specified
[109]	TOTEM Framework (Token for Controlled Computation), Blockchain, Machine Learning	Not specified	Data analysis without moving data, Data security, Privacy of prosumers	TOTEM framework for analyzing data without moving it, Ensuring data security and privacy	Not specified
[110]	Prophet Model, ARIMA	Two datasets	Demand response management, Forecasting	Effectiveness of Prophet model in demand response management	Various evaluation metrics
[111]	CNN with RUSBoost MRFO and RUSBoost BSA models	Not specified	Electric power theft detection	Accuracy of proposed approaches: RUS-MRFO: 91.5 %, RUS-BSA: 93.5 %	Accuracy percentages
[112]	GA-LSTM (Genetic Algorithm - Long Short Term Memory)	Pennsylvania-New Jersey-Maryland Interconnection (PJM) energy consumption data	Genetic Algorithm, Long Short Term Memory	Better performance compared to existing benchmarks	Various performance evaluation metrics
[113]	Optimal Singular Value Decomposition (SVD), Intelligent Optimization Methods	Wide range of data types	Data compression, Retrieval quality, Compression ratio	Compression level dominates existing SVD rank reduction methods	Compression level
[114]	Deep Siamese Network (DSN), Adaptive Synthesis (ADASYN)	Real-time smart meters' data	Imbalanced class problem, Feature extraction	Effective in resolving imbalanced class problem	AUC, F1-Score, Precision, Recall
[115]	IoT, Cloud Computing	IMWSNs measurements during events monitoring and control	Channel detection, Channel assignment, Packets forwarding	Useful for designing algorithms for real-time events monitoring and control	Not specified
[116]	Fuzzy Logic, Genetic Algorithm	Not specified	Fuzzy logic, Genetic algorithm	Optimal values for maximizing profit and predicting future power demands	Not specified
[117]	Fog Computing, Cloud Computing	Exemplary SG network	Planning and Placement of Fog computing in smart Grid (PPFG)	Optimization of FCN location, capacity, and number	Response delay, Energy consumption
[118]	Differentially Private Clustering, Infinite Gaussian Mixture Model (IGMM)	Not specified	Privacy-preserving cluster analysis	Privacy-preserving, Efficient	Security analysis, Performance evaluation
[119]	Hybrid Outlier Rejection Methodology (HORM)	Not specified	Outlier rejection for load forecasting	Outperforms recent methods in terms of accuracy, precision, recall, F1-measure	Accuracy, Precision, Recall, F1-measure
[120]	Edge Computing, Big Data Analytics	Not specified	Non-Technical Loss fraud detection	Efficient detection of non-technical loss frauds	Detection speed
[121]	Hadoop-Hbase, Cassandra, Elasticsearch, MongoDB	Large scale smart grid data generator	Data analysis techniques	Performance benchmark for different frameworks	Not specified
[122]	Big Data Anonymous Batch Verification, Edge Computing, Certificateless Aggregate Signature (CL-AS)	Not specified	Batch-verifiable authentication, Privacy preservation	Efficient detection of power injection without exposing private information	Not specified
[123]	Hybrid MDM/R Architecture, MapReduce, Massive Parallel Processing Database	Not specified	Scheduling of workloads, Energy consumption, RE integration	Modification of supply-following algorithm, New hybrid architecture	Not specified
[124]	Temporal, Functional, Spatial Big Data Computing	Large-scale smart grid	Data extraction, Computing efficiency	Promising computing efficiency, Bandwidth savings	Convergence ratio, Improvement ratio
[125]	Fiber-Wireless (FiWi) Enhanced Smart Grid	Not specified	Data acquisition under failures	Constrained optimization problem, OERA, GARA, HGRA	Not specified
[126]	Cloud Computing, Grid Computing	Not specified	Information management, Data storage	Business intelligence architecture, Multivariable, Multi-dimensional analysis	Not specified
[127]	Fog Computing, Cloud Computing	Not specified	Electrical load forecasting	3-tiers architecture, Data pre-processing, Load prediction	Precision, Recall, Accuracy, F-measure
[128]	Random Forest, Relief-F, Grey Correlation Analysis, Kernel PCA, Differential Evolution, Support Vector Machine	Not specified	Feature selection, Feature extraction, Price forecasting	Hybrid feature selector, Dimensionality reduction, DE-SVM classifier	Price forecasting Performance
[129]	Cloud Computing, Power Big Data	Not specified	Power efficiency analysis	Business intelligence architecture, Multivariable, Multi-dimensional analysis	Not specified

Table 4
Accuracy comparison of big data analytics in smart grids.

Study	Methodology	Accuracy
[111]	Novel CNN with RUSBoost MRFO and RUSBoost BSA models. Accuracy: rus-MRFO 91.5 %, rus-BSA 93.5 %. SVM accuracy 71 %, LR 63 %. Smote algorithm for balancing. Boosting techniques outperform.	SVM: 71 %, LR: 63 %, CNN: 85.1 %, rus-MRFO: 90 %, rus-BSA: 93 %
[114]	ADASYN handles imbalance. CNN-LSTM integrated deep siamese network. High AUC, mAP, precision, recall, MaP, accuracy, and F1-Score. Maintains performance for different training ratios. Simulation results validate effectiveness.	SDN: AUC 0.93 %, mAP 0.9 %, Accuracy (60 %: 0.839, 70 %: 0.844, 80 %: 0.953)
[119]	ELF strategy with DP2 and LP2 phases. Proposed HORM for outlier rejection, outperforms recent methods. Experimental results show higher accuracy, precision, recall, and F1-measure.	'All Data': 87 %, HORM: 95 %
[127]	ELF strategy with DP2 and LP2 phases. FBFS for feature selection. Experimental results show improved efficiency in terms of precision, recall, accuracy, and F-measure.	FSSO2: 0.80, FSSO4: 0.93
[106]	Comparison of linear regression, logistic regression, and K-Nearest Neighbors. Linear regression gives highest accuracy (98 %).	Linear Regression: 98 %, Logistic Regression: 96 %, K-NN: 92 %
[112]	GA-LSTM method based on Genetic Algorithm. GA-LSTM outperforms existing benchmarks. Multi-threaded environment for increased convergence speed. Results validated on PJM energy consumption data.	GA: 82.42 (daily), 80.27 (weekly), Random Approach: 51.26 (daily), 48.22 (weekly)
[124]	Novel framework for large-scale smart grid. Functional and spatial dimensions considered. Promising computing efficiency (95 % convergence ratio) and bandwidth savings (81 % improvement ratio over benchmarks).	Computing Efficiency: 95 % convergence ratio, Bandwidth: 81 % improvement ratio
[128]	Hybrid feature selector (RF + Relief-F + GCA), KPCA for dimensionality reduction, DE-based SVM classifier. Superior performance compared to other methods.	HSEC: >90 %, Frameworks A, B, C, and HSEC show improvement in accuracy

Rabie et al. [127] and Nivedha et al. [126] emphasized the need for a comprehensive data privacy framework, recognizing the potential risks that ongoing data collection via smart meters and sensors may pose to sensitive consumer information. Rabie et al. [127] highlighted the importance of implementing a fog-based load forecasting approach by proposing a three-tier structure that improves the acquisition, processing, and retention of smart metre data prior to its transmission to the cloud. Due to the sensitive nature of the data involved, this strategy requires meticulous deliberation regarding privacy issues. This is consistent with the wider difficulties presented by data privacy within the smart grid domain.

In light of the obstacles encountered, Guan et al. [118] proposed that differential privacy techniques be integrated into cluster analysis as a means of safeguarding the confidentiality of smart grid data. The algorithm they suggest, IDPC, integrates differential privacy with nonparametric Bayesian techniques, providing a solution that takes into account the dynamic characteristics of clustering while maintaining privacy. Furthermore, the research conducted by Han and Xiao [120] presented the notion of edge computing-enabled non-technical loss fraud detection, underscoring the criticality of safeguarding big data analytics at the periphery in order to avert fraudulent activities. These proposed solutions are in line with the wider requirement for secure edge computing infrastructure and privacy-preserving analytics. They underscore the importance of sophisticated frameworks and techniques in mitigating security issues related to large-scale data in smart grids. It is critical to confront encryption obstacles, as demonstrated by research such as the one proposed by Kamil and Ogundoyin [122] for an anonymous bulk verification scheme for big data. By integrating highly efficient certificateless aggregate signature algorithms, their methodology guarantees the authentication of power offers transmitted through vehicular networks and 5G smart grid slices in a secure manner.

RQ3. How do machine learning and artificial intelligence techniques enhance decision-making processes in smart grid applications?

Machine Learning and Artificial Intelligence are instrumental in transforming the decision-making process. The challenges presented by the expanding intricacy of data produced by smart grids are motivating the growing dependence on these technologies. This is evident in research endeavors like Jiang et al.'s [105] investigation into the application of Canonical Correlation Analysis in comprehending electricity consumption patterns. Integration of machine learning and artificial intelligence techniques was motivated by the need to resolve issues pertaining to the manipulation and analysis of large datasets, the transmission of data efficiently, and the utilization of diverse data sources effectively.

Load forecasting has exhibited potential for machine learning techniques, specifically supervised learning, as Gupta and Chaturvedi [106] illustrated in their study on Adaptive Energy Management. The study by Park and Huh [108] on Apache Kafka serves as an illustration of how unsupervised learning enhances the efficiency of data collection and utilization in distributed manufacturing networks. Centralized systems, such as ESB or EAI, which are used for data collection in manufacturing traditionally, have problems integrating data from different networks in factories. Data collection across distributed or autonomous networks is inefficient and vulnerable to security breaches due to this centralized approach. The goal of implementing Kafka was to facilitate user analysis and utilization of data across various manufacturing networks. In addition, research by Kumari et al.'s [110] approach to Demand Response Management emphasized how Reinforcement Learning optimizes decision-making processes through the acquisition of knowledge from interactions in dynamic environments.

Artificial intelligence techniques, including Knowledge-Based Systems for grid optimization and Expert Systems for defect detection, improve decision-making. The incorporation of rule-based methodologies and machine learning, as demonstrated in the research conducted by Jose et al. [109] concerning the integration of blockchain and big data, enhances the security and reliability of energy management. The insights provided by these artificial intelligence techniques are crucial for ensuring the security and dependability of smart grid operations.

It is evident that machine learning and artificial intelligence work in tandem in ensemble learning and hybrid models. The optimization SVD method proposed by Hashemipour et al. [113] illustrates how integrating distinct models improves both data compression and analysis. An instance of ensemble learning strategies in action is the one examined by Wang et al. [128] in their investigation of electricity price forecasting. A novel model for forecasting electricity prices was created to overcome the difficulties in managing large amounts of price data in smart grids. It combined three modules: feature extraction using a combination of Principle Component Analysis and Kernel function for dimensionality reduction, a hybrid feature selector utilizing Random Forest and Relief-F algorithm based on Grey Correlation Analysis, and a price classification forecast using a Support Vector Machine classifier based on differential evolution. The suggested model outperformed other approaches, as evidenced by numerical results, and this made it an efficient solution for smart grids' efficient forecasting of electricity prices.

When considering issues related to data quality, privacy, interoperability, and integration, these studies emphasize the need for a holistic strategy when integrating machine learning and AI. Practical

applications of real-world case studies, such as the research conducted by Rabie et al. [119] regarding anomalous rejection methodologies, offer valuable insights into surmounting obstacles and attaining efficient decision-making within the ever-evolving domain of smart grid technologies. The integration of machine learning and artificial intelligence into smart grids is anticipated to yield advantages that will propel progress in grid resilience and significantly influence the trajectory of energy management in the coming years.

RQ4. What role does blockchain technology play in ensuring secure and transparent transactions within the context of energy smart grids?

Numerous studies provide evidence that blockchain technology has emerged as a revolutionary influence in safeguarding the confidentiality and visibility of transactions within energy smart grids. The significance of blockchain technology in establishing a decentralized and immutable ledger to safeguard energy transactions was emphasized by Jiang et al. [105]. The blockchain employs cryptographic methods to bolster security measures, thereby ensuring that unauthorized modifications are thwarted. The integration of cryptographic techniques with this decentralized and secure ledger enhances the dependability of transactions, which is consistent with the conclusions drawn by Hashemipour et al. [113]. Highlighting the significance of data integrity, their research illustrates the most effective utilization of SVD for the compression of large amounts of data within smart grids. The cryptographic mechanisms of the blockchain are consistent with these discoveries, thereby enhancing security.

The incorporation of smart contracts into blockchain technology enhances the automation and security of transactions, thereby diminishing the necessity for intermediaries, as evidenced by research conducted by Jose et al. [109] and Akram et al. [111]. Jose et al. [109] underscored the importance of smart contracts in the secure management of data and the assurance of reliable transactions. In a similar vein, the efficacy of smart contracts supported by blockchain technology in enhancing the precision of prediction analysis for large-scale data within smart grids is illustrated by Akram et al. [111]. Because electric power theft has proven difficult to combat with Advanced Metering Infrastructure (AMI), researchers are looking to machine learning algorithms for solutions. The significance of blockchain in automating and securing transactions, thereby reducing the hazards linked to human involvement, is highlighted by these results.

It has been demonstrated that decentralized energy trading platforms, enabled by blockchain technology, improve peer-to-peer transactions while reducing the influence of centralized authorities. This is consistent with the findings of Kamil and Ogundoyin's [122] research, in which the authors advocate for an anonymous bulk verification scheme for big data, with an emphasis on the significance of safeguarding the privacy of users. To protect user privacy while authenticating large power bids from electric vehicles (EVs) on vehicular networks and 5G smart grid slices, a novel method utilizing the certificateless aggregate signature (CL-AS) algorithm is suggested. While maintaining participant privacy, this approach guarantees batch-verifiable authentication, with security reliant on the Discrete Logarithm Problem's (DLP) intractability. Performance analyses demonstrated that it is more efficient than comparable schemes, indicating improved acceptability and fairness in the energy market.

7. Implications for future research

Despite the increasing amount of research on smart grids and big data analytics, many issues still need to be addressed before this technology can be widely used. One of the main obstacles to data exchange and access is the lack of standard formats and system compatibility. Most studies are conducted with ideal data conditions, which can limit progress in research due to privacy concerns. The complexity of smart grid analytics makes it imperative that specialists from different domains work together interdisciplinary. Only then can its full potential be realized. Given these challenges, it is clear that smart grid research needs to focus on software and hardware availability, interoperability, data privacy, regulatory compliance, computing, scalability, and optimization algorithms if we want them to work effectively and last. Improving grid resilience and dependability requires leveraging heterogeneous data sources, fusing analytics with real-time control, and developing computational algorithms.

Several important areas are covered by the suggestions for future research that aim to improve the efficiency and real-world application of data science advancements in smart environment technology. Prioritizing privacy and security assessments, we will delve into advanced encryption methods and secure computation techniques to safeguard sensitive consumer data while we extract meaningful insights. Other areas of focus will include integrating diverse data analytics techniques to explore the potential of hybrid models and fostering collaborations with smart grid operators for real-world implementation studies.

Contributing to a thorough understanding and successful implementation of big data innovations in smart grids while considering social and environmental impacts, this project investigates the integration of blockchain technology to improve data integrity and transparency, builds decision support systems and user-friendly interfaces to help stakeholders understand and use analytics insights, investigates interoperability standards, assesses analytics solutions' resilience, and conducts cost-benefit analyses.

The field of smart environment technologies presents a plethora of real-world applications that might revolutionize the data science field. Demand response management is one of the most common applications. Here, state-of-the-art data analytics methods such as genetic algorithms and machine learning may be used to better forecast and manage energy demand, maximizing the supply-demand balance in smart grids. Machine learning techniques like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) can be used in conjunction with big data analytics to develop dependable systems that detect and prevent power theft. Another practical use is the development of intricate systems for the detection of fraudulent activity. Fog computing's incorporation into smart grids offers practical answers to problems with latency and energy consumption. The network's intentional deployment of fog computing nodes allows for real-time processing and decision-making. This results in much shorter reaction times and improved performance for smart grid applications.

8. Conclusion

The present systematic review investigates the intersection of advancements in data science and smart environment technologies, with a particular focus on smart utilities. The objective is to clarify significant developments, obstacles, and consequences associated with the incorporation of big data analytics in influencing the trajectory of energy management. With the growing need for energy solutions that are both efficient and sustainable, the convergence of data science and smart infrastructure assumes a more critical role.

This exhaustive analysis of 25 research papers focuses on the convergence of big data analytics and smart grids, offering insights into the various applications and challenges that exist within this domain. The research papers present successful methodologies for canonical correlation analysis and adaptive energy management, as well as novel platforms, techniques for responding to demand, and approaches for achieving optimal compression. The significance of big data analytics in tackling critical challenges in smart grids is underscored by inquiries into load forecasting, energy efficiency, electricity theft detection, and predictive modeling. These investigations emphasize the contributions of big data analytics in areas such as energy efficiency, privacy protection, and predictive modeling.

The implications of these findings extend to future energy landscapes. The application of machine learning algorithms to demand response management and larceny detection studies demonstrates the

potential for enhanced security measures. Future investigations into smart grids should place emphasis on the improvement of machine learning models, the advancement of predictive analytics, and the enhancement of data processing efficiency. This all-encompassing synopsis establishes the incorporation of big data analytics as a paradigm shifter in tackling obstacles and guiding intelligent power systems towards a future characterized by enhanced resilience and adaptability.

To improve the accuracy of predictive analytics in smart grid management, future research should concentrate on developing advanced machine learning algorithms. These algorithms should include deep learning neural networks as well as reinforcement learning. Investigating distributed and edge computing architectures for processing data in real-time is necessary to address scalability issues. It is necessary to conduct research into blockchain-based frameworks for secure data sharing and transaction verification to integrate blockchain technology and enhance data security and decentralization. Smart energy systems that are more autonomous and resilient can be developed by researchers by taking advantage of these innovations.

CRediT authorship contribution statement

Hamed Taherdoost: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] N. Roztocki, P. Soja, H.R. Weistroffer, The Role of Information and Communication Technologies in Socioeconomic Development: towards a Multi-Dimensional Framework, Taylor & Francis, 2019, pp. 171–183.
[2] A. Harerimana, N.G. Mtshali, Types of ICT applications used and the skills' level of nursing students in higher education: a cross-sectional survey, International Journal of Africa Nursing Sciences 11 (2019) 100163.
[3] O.I. Olatoye, F. Nekhwevha, N. Muchaonyerwa, ICT literacy skills proficiency and experience on the use of electronic resources amongst undergraduate students in selected Eastern Cape Universities, South Africa, Libr. Manag. 42 (6/7) (2021) 471–479.
[4] I. Harris, Y. Wang, H. Wang, ICT in multimodal transport and technological trends: unleashing potential for the future, Int. J. Prod. Econ. 159 (2015) 88–103.
[5] B. Gera, et al., Leveraging AI-enabled 6G-driven IoT for sustainable smart cities, Int. J. Commun. Syst. 36 (16) (2023) e5588.
[6] D. Soldani, S.A. Illingworth, 5G AI-Enabled Automation, Elsevier, 2020.
[7] G. Elia, et al., A multi-dimension framework for value creation through big data, Ind. Market. Manag. 90 (2020) 617–632.
[8] A. Urbinati, et al., Creating and capturing value from Big Data: a multiple-case study analysis of provider companies, Technovation 84 (2019) 21–36.
[9] T.J. Saleem, M.A. Chishti, Data analytics in the internet of things: a survey, Scalable Comput. Pract. Exp. 20 (4) (2019) 607–630.
[10] P. Sunhare, R.R. Chowdhary, M.K. Chattopadhyay, Internet of things and data mining: an application oriented survey, Journal of King Saud University-Computer and Information Sciences 34 (6) (2022) 3569–3590.
[11] E. Badidi, Z. Mahrez, E. Sabir, Fog computing for smart cities' big data management and analytics: a review, Future Internet 12 (11) (2020) 190.
[12] K. Soomro, et al., Smart city big data analytics: an advanced review, Wiley Interdisciplinary Reviews: Data Min. Knowl. Discov. 9 (5) (2019) e1319.
[13] A. Jindal, N. Kumar, M. Singh, A unified framework for big data acquisition, storage, and analytics for demand response management in smart cities, Future Generat. Comput. Syst. 108 (2020) 921–934.
[14] A. Kumari, et al., When blockchain meets smart grid: secure energy trading in demand response management, IEEE Network 34 (5) (2020) 299–305.
[15] I.H. Sarker, Smart City Data Science: towards data-driven smart cities with open research issues, Internet of Things 19 (2022) 100528.
[16] A. Ullah, et al., Smart cities: the role of Internet of Things and machine learning in realizing a data-centric smart environment, Complex & Intelligent Systems (2023) 1–31.
[17] V. Grossi, et al., Data science: a game changer for science and innovation, International Journal of Data Science and Analytics 11 (2021) 263–278.
[18] S.B. Aitallah, et al., Leveraging Deep Learning and IoT big data analytics to support the smart cities development: review and future directions, Computer Science Review 38 (2020) 100303.
[19] R. Vinuesa, et al., The role of artificial intelligence in achieving the Sustainable Development Goals, Nat. Commun. 11 (1) (2020) 1–10.
[20] W. Leal Filho, et al., Using data science for sustainable development in higher education, Sustain. Dev. 32 (1) (2024) 15–28.
[21] V. Sebestyén, T. Czvetkó, J. Abonyi, The applicability of big data in climate change research: the importance of system of systems thinking, Front. Environ. Sci. 9 (2021) 70.
[22] R. Dorschel, Discovering needs for digital capitalism: the hybrid profession of data science, Big Data & Society 8 (2) (2021) 20539517211040760.
[23] S.E. Bibri, Data-driven smart sustainable cities of the future: an evidence synthesis approach to a comprehensive state-of-the-art literature review, Sustainable Futures 3 (2021) 100047.
[24] G. Amato, et al., How data mining and machine learning evolved from relational data base to data science, A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years (2018) 287–306.
[25] V. Grossi, et al., Data science at SoBigData: the European research infrastructure for social mining and big data analytics, International Journal of Data Science and Analytics 6 (2018) 205–216.
[26] A. Kumari, et al., Blockchain-driven real-time incentive approach for energy management system, Mathematics 11 (4) (2023) 928.
[27] I. Martinez, E. Viles, I.G. Olaizola, Data science methodologies: current challenges and future approaches, Big Data Research 24 (2021) 100183.
[28] G. Nandi, R.K. Sharma, Data Science Fundamentals and Practical Approaches: Understand Why Data Science Is the Next, BPB Publications, 2020.
[29] S. Molin, K. Jee, Hands-On Data Analysis with Pandas: A Python Data Science Handbook for Data Collection, Wrangling, Analysis, and Visualization, Packt Publishing Ltd, 2021.
[30] M. Birjali, M. Kasri, A. Beni-Hssane, A comprehensive survey on sentiment analysis: approaches, challenges and trends, Knowl. Base Syst. 226 (2021) 107134.
[31] I.H. Sarker, Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective, SN Computer Science 2 (5) (2021) 377.
[32] P. Mikalef, R. van de Wetering, J. Krogstie, Building dynamic capabilities by leveraging big data analytics: the role of organizational inertia, Inf. Manag. 58 (6) (2021) 103412.
[33] A. Nielsen, Practical Time Series Analysis: Prediction with Statistics and Machine Learning, O'Reilly Media, 2019.
[34] R. Agarwal, V. Dhar, Big Data, Data Science, and Analytics: the Opportunity and Challenge for IS Research, INFORMS, 2014, pp. 443–448.
[35] K.M. Tolle, D.S.W. Tansley, A.J. Hey, The fourth paradigm: data-intensive scientific discovery [point of view], Proc. IEEE 99 (8) (2011) 1334–1337.
[36] A. Elragal, R. Klischewski, Theory-driven or process-driven prediction? Epistemological challenges of big data analytics, Journal of Big Data 4 (2017) 1–20.
[37] M. Frické, Big data and its epistemology, Journal of the association for information science and technology 66 (4) (2015) 651–661.
[38] R. Kitchin, Big Data, new epistemologies and paradigm shifts. Big Data & Society 1 (1) (2014) 1–12 (Sage Publications).
[39] O. Müller, et al., Utilizing big data analytics for information systems research: challenges, promises and guidelines, Eur. J. Inf. Syst. 25 (2016) 289–302.
[40] F. Tao, et al., Data-driven smart manufacturing, J. Manuf. Syst. 48 (2018) 157–169.
[41] Y. Gao, et al., A review on recent advances in vision-based defect recognition towards industrial intelligence, J. Manuf. Syst. 62 (2022) 753–766.
[42] R. Sahal, J.G. Breslin, M.I. Ali, Big data and stream processing platforms for Industry 4.0 requirements mapping for a predictive maintenance use case, J. Manuf. Syst. 54 (2020) 138–151.
[43] N. Tuptuk, S. Hailes, Security of smart manufacturing systems, J. Manuf. Syst. 47 (2018) 93–106.
[44] J. Wang, et al., Deep learning for smart manufacturing: methods and applications, J. Manuf. Syst. 48 (2018) 144–156.
[45] S. Kergroach, Industry 4.0: new challenges and opportunities for the labour market, Форсайт 11 (2017) 6–8, 4 (eng).
[46] A. Kumari, S. Tanwar ρreveal, An ai-based big data analytics scheme for energy price prediction and load reduction, in: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2021.
[47] F. Pallonetto, M. De Rosa, D.P. Finn, Impact of intelligent control algorithms on demand response flexibility and thermal comfort in a smart grid ready residential building, Smart Energy 2 (2021) 100017.
[48] M. Babar, M.U. Tariq, M.A. Jan, Secure and resilient demand side management engine using machine learning for IoT-enabled smart grid, Sustain. Cities Soc. 62 (2020) 102370.
[49] P.-A. Langendahl, et al., Smoothing peaks and troughs: intermediary practices to promote demand side response in smart grids, Energy Res. Social Sci. 58 (2019) 101277.

[50] Z. Wang, et al., Incentive based emergency demand response effectively reduces peak load during heatwave without harm to vulnerable groups, Nat. Commun. 14 (1) (2023) 6202.
[51] A. Ucar, M. Karakose, N. Kırımça, Artificial intelligence for predictive maintenance applications: key components, trustworthiness, and future trends, Appl. Sci. 14 (2) (2024) 898.
[52] P. Coandă, M. Avram, V. Constantin, A state of the art of predictive maintenance techniques, in: IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2020.
[53] J. Passlick, et al., Predictive maintenance as an internet of things enabled business model: a taxonomy, Electron. Mark. 31 (2021) 67–87.
[54] M.J.B. Kabeyi, O.A. Olanrewaju, Smart grid technologies and application in the sustainable energy transition: a review, Int. J. Sustain. Energy 42 (1) (2023) 685–758.
[55] F.G. Gonzalez, An intelligent controller for the smart grid, Procedia Comput. Sci. 16 (2013) 776–785.
[56] F. Garzia, et al., Meeting user needs through building automation and control systems: a review of impacts and benefits in office environments, Buildings 13 (10) (2023) 2530.
[57] A. Kumari, et al., Blockchain-based peer-to-peer transactive energy management scheme for smart grid system, Sensors 22 (13) (2022) 4826.
[58] Y. Zhang, T. Huang, E.F. Bompard, Big data analytics in smart grids: a review, Energy Informatics 1 (1) (2018).
[59] A. Kumari, et al., ET-DeaL: a P2P smart contract-based secure energy trading scheme for smart grid systems, in: IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), IEEE, 2020.
[60] E.J. Nielsen, B. Diskin, High-performance aerodynamic computations for aerospace applications, Parallel Comput. 64 (2017) 20–32.
[61] S. Meng, X. He, X. Tian, Research on Fintech development issues based on embedded cloud computing and big data analysis, Microprocess. Microsyst. 83 (2021) 103977.
[62] W.A. De Jong, et al., Utilizing high performance computing for chemistry: parallel computational chemistry, Phys. Chem. Chem. Phys. 12 (26) (2010) 6896–6920.
[63] K. Sanbonmatsu, C.-S. Tung, High performance computing in biology: multimillion atom simulations of nanoscale systems, J. Struct. Biol. 157 (3) (2007) 470–480.
[64] P. Angelikopoulos, C. Papadimitriou, P. Koumoutsakos, Bayesian uncertainty quantification and propagation in molecular dynamics simulations: a high performance computing framework, J. Chem. Phys. 137 (14) (2012).
[65] D. Sarojini, et al., Towards developing a guide to choosing national high-performance computing resources, in: Practice and Experience in Advanced Research Computing, 2023, pp. 382–385.
[66] S. Zhu, et al., Intelligent computing: the latest advances, challenges, and future, Intelligent Computing 2 (2023) 6.
[67] K. Volovich, S. Denisov, The main scientific and technical problems of using hybrid HPC clusters in materials science, Russ. Microelectron. 49 (8) (2020) 574–579.
[68] A.J. Banegas-Luna, et al., Advances in distributed computing with modern drug discovery, Expet Opin. Drug Discov. 14 (1) (2019) 9–22.
[69] M. Rahimi-Gorji, et al., Optimization of intraperitoneal aerosolized drug delivery using computational fluid dynamics (CFD) modeling, Sci. Rep. 12 (1) (2022) 6305.
[70] S. Usman, et al., Data locality in high performance computing, big data, and converged systems: an analysis of the cutting edge and a future system architecture, Electronics 12 (1) (2022) 53.
[71] S.A. Zenios, High-performance computing in finance: the last 10 years and the next, Parallel Comput. 25 (13–14) (1999) 2149–2175.
[72] P. Raj, et al., Real-time analytics using high-performance computing, High-Performance Big-Data Analytics: Computing Systems and Approaches (2015) 161–185.
[73] C. Barakat, et al., Lessons learned on using high-performance computing and data science methods towards understanding the acute respiratory distress syndrome (ARDS), in: 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology, MIPRO 2022 - Proceedings, 2022.
[74] G. Mellone, et al., Democratizing the computational environmental marine data science: using the High-Performance Cloud-Native Computing for inert transport and diffusion Lagrangian modelling, in: 2022 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters, MetroSea 2022 - Proceedings, 2022.
[75] A. Oujja, et al., High-performance computing for SARS-CoV-2 RNAs clustering: a data science-based genomics approach, Genomics and Informatics 19 (4) (2021).
[76] J.P. Courneya, A. Mayo, High-performance computing service for bioinformatics and data science, J. Med. Libr. Assoc. 106 (4) (2018) 494–495.
[77] S.D. Belov, et al., High-performance computing platforms for organizing the educational process on the basis of the international school "data science", in: CEUR Workshop Proceedings, 2019.
[78] H.J. Siegel, S. Suryanarayanan, Plenary panel: convergence of high-performance computing and communication, smart city, and data sciences and systems: fields helping grand challenges and each other, in: Proceedings - 2017 IEEE 19th Intl Conference on High Performance Computing and Communications, HPCC 2017, 2017 IEEE 15th Intl Conference on Smart City, SmartCity 2017 and 2017 IEEE 3rd Intl Conference on Data Science and Systems, DSS 2017, 2017.
[79] S. Goto, D.K. McGuire, S. Goto, The future role of high-performance computing in cardiovascular medicine and science -impact of multi-dimensional data analysis, J. Atherosclerosis Thromb. 29 (5) (2022) 559–562.
[80] B. Prashanth, et al., Optimization factors with high performance computing and data science based implementations with metaheuristics, AIP Conf. Proc. 2418 (1) (2022) 020043.
[81] M. Riedel, et al., Practice and experience using high performance computing and quantum computing to speed-up data science methods in scientific applications, in: 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology, MIPRO 2022 - Proceedings, 2022.
[82] S. Dash, et al., Big data in healthcare: management, analysis and future prospects, Journal of big data 6 (1) (2019) 1–25.
[83] U. Sivarajah, et al., Critical analysis of Big Data challenges and analytical methods, J. Bus. Res. 70 (2017) 263–286.
[84] M.M. Rathore, et al., Real-time secure communication for smart city in high-speed big data environment, Future Generat. Comput. Syst. 83 (2018) 638–652.
[85] A. Celesti, M. Fazio, A framework for real time end to end monitoring and big data oriented management of smart environments, J. Parallel Distr. Comput. 132 (2019) 262–273.
[86] H. Yu, Z. Yang, R.O. Sinnott, Decentralized big data auditing for smart city environments leveraging blockchain technology, IEEE Access 7 (2019) 6288–6296.
[87] A. Xu, W. Zeng, Dynamic optimization modeling of smart tourism information system using VRGIS in big data environment, Comput. Intell. Neurosci. 2022 (2022).
[88] M. Babar, F. Arif, Real-time data processing scheme using big data analytics in internet of things based smart transportation environment, J. Ambient Intell. Hum. Comput. 10 (10) (2019) 4167–4177.
[89] H. Chen, et al., An efficient recommendation filter model on smart home big data analytics for enhanced living environments, Sensors 16 (10) (2016).
[90] G. Li, et al., A partial order reduction based method for big data preprocessing in smart grid environment, Dianli Xitong Zidonghua/Automation of Electric Power Systems 40 (7) (2016) 98–106.
[91] Z. Elaggoune, R. Maamri, I. Boussebough, A fuzzy agent approach for smart data extraction in big data environments, Journal of King Saud University - Computer and Information Sciences 32 (4) (2020) 465–478.
[92] M. Sun, J. Zhang, Research on the application of block chain big data platform in the construction of new smart city for low carbon emission and green environment, Comput. Commun. 149 (2020) 332–342.
[93] S. Nagarajan, K. Perumal, Structured and unstructured information extraction using text mining and natural language processing techniques, International Journal on Recent and Innovation Trends in Computing and Communication 5 (11) (2020) 32–43.
[94] K. Adnan, R. Akbar, Limitations of information extraction methods and techniques for heterogeneous unstructured big data, Int. J. Eng. Bus. Manag. 11 (2019) 1847979019890771.
[95] M.-E. Vidal, S. Jozashoori, A. Sakor, Semantic data integration techniques for transforming big biomedical data into actionable knowledge, in: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2019.
[96] A. Thomas, S. Sangeetha, An innovative hybrid approach for extracting named entities from unstructured text data, Comput. Intell. 35 (4) (2019) 799–826.
[97] P.R. da Silva, et al., Extraction of Useful Information from Unstructured Data in Software Engineering: A Systematic Mapping, CIbSE, 2020.
[98] F. Benites, Information retrieval and knowledge extraction for academic writing, in: Digital Writing Technologies in Higher Education: Theory, Research, and Practice, Springer, 2023, pp. 303–315.
[99] A. Holzinger, Introduction to Machine Learning & Knowledge Extraction (Make), Multidisciplinary Digital Publishing Institute, 2019, pp. 1–20.
[100] M. Mameli, et al., Deep learning approaches for fashion knowledge extraction from social media: a review, IEEE Access 10 (2021) 1545–1576.
[101] A. Mahani, A.R. Baba-Ali, A new rule-based knowledge extraction approach for imbalanced datasets, Knowl. Inf. Syst. 61 (2019) 1303–1329.
[102] L.-L. Xie, et al., Knowledge extraction for solving resource-constrained project scheduling problem through decision tree, Eng. Construct. Architect. Manag. (2023). Vol. ahead-of-print No. ahead-of-print, https://www.emerald.com/insight/content/doi/10.1108/ECAM-04-2022-0345/full/html.
[103] D. Cabrera, et al., Knowledge extraction from deep convolutional neural networks applied to cyclo-stationary time-series classification, Inf. Sci. 524 (2020) 1–14.
[104] H. Taherdoost, M. Madanchian, Artificial intelligence and sentiment analysis: a review in competitive research, Computers 12 (2) (2023) 37.
[105] Z. Jiang, et al., Canonical correlation analysis and visualization for big data in smart grid, IEEE Journal on Emerging and Selected Topics in Circuits and Systems 13 (3) (2023) 702–711.
[106] R. Gupta, K.T. Chaturvedi, Adaptive energy management of big data analytics in smart grids, Energies 16 (16) (2023).
[107] J. Wang, et al., A post-evaluation system for smart grids based on microservice framework and big data analysis, Electronics 12 (7) (2023).
[108] S. Park, J.H. Huh, A study on big data collecting and utilizing smart factory based grid networking big data using Apache Kafka, IEEE Access 11 (2023) 96131–96142.
[109] D.T. Jose, et al., Integrating big data and blockchain to manage energy smart grids—TOTEM framework, Blockchain: Research and Applications 3 (3) (2022).
[110] S. Kumari, N. Kumar, P.S. Rana, A big data approach for demand response management in smart grid using the Prophet model, Electronics 11 (14) (2022).
[111] R. Akram, et al., Towards big data electricity theft detection based on improved rusboost classifiers in smart grid, Energies 14 (23) (2021).

[112] S. Kumari, N. Kumar, P.S. Rana, Big data analytics for energy consumption prediction in smart grid using genetic algorithm and long short term memory, Comput. Inf. 40 (1) (2021) 29–56.
[113] S. Hashemipour, et al., Optimal singular value decomposition based big data compression approach in smart grids, IEEE Trans. Ind. Appl. 57 (4) (2021) 3296–3305.
[114] N. Javaid, N. Jan, M.U. Javed, An adaptive synthesis to handle imbalanced big data with deep siamese network for electricity theft detection in smart grids, J. Parallel Distr. Comput. 153 (2021) 44–52.
[115] M. Faheem, et al., Big Data acquired by Internet of Things-enabled industrial multichannel wireless sensors networks for active monitoring and control in the smart grid Industry 4.0, Data Brief 35 (2021).
[116] S.M. Je, J.H. Huh, Estimation of future power consumption level in smart grid: application of fuzzy logic and genetic algorithm on big data platform, Int. J. Commun. Syst. 34 (2) (2021).
[117] M.M. Hussain, M.M.S. Beg, M.S. Alam, Fog computing for big data analytics in IoT aided smart grid networks, Wireless Pers. Commun. 114 (4) (2020) 3395–3418.
[118] Z. Guan, et al., A differentially private big data nonparametric bayesian clustering algorithm in smart grid, IEEE Transactions on Network Science and Engineering 7 (4) (2020) 2631–2641.
[119] A.H. Rabie, et al., A new outlier rejection methodology for supporting load forecasting in smart grids based on big data, Cluster Comput. 23 (2) (2020) 509–535.
[120] W. Han, Y. Xiao, Edge computing enabled non-technical loss fraud detection for big data security analytic in Smart Grid, J. Ambient Intell. Hum. Comput. 11 (4) (2020) 1697–1708.
[121] M.H. Ansari, V. Tabatab Vakili, B. Bahrak, Evaluation of big data frameworks for analysis of smart grids, Journal of Big Data 6 (1) (2019).
[122] I.A. Kamil, S.O. Ogundoyin, A big data anonymous batch verification scheme with conditional privacy preservation for power injection over vehicular network and 5G smart grid slice, Sustainable Energy, Grids and Networks 20 (2019).
[123] A. Mehenni, et al., An optimal big data processing for smart grid based on hybrid MDM/R architecture to strengthening RE integration and EE in datacenter, J. Ambient Intell. Hum. Comput. 10 (9) (2019) 3709–3722.
[124] W. Hou, et al., Temporal, functional and spatial big data computing framework for large-scale smart grid, IEEE Transactions on Emerging Topics in Computing 7 (3) (2019) 369–379.
[125] H. Guo, J. Liu, L. Zhao, Big data acquisition under failures in FiWi enhanced smart grid, IEEE Transactions on Emerging Topics in Computing 7 (3) (2019) 420–432.
[126] R. Nivedha, S. Arshiya Sulthana, A secure cloud computing based framework for big data information management of smart grid, Int. J. Innovative Technol. Explor. Eng. 8 (6 Special Issue 4) (2019) 1235–1238.
[127] A.H. Rabie, et al., A fog based load forecasting strategy for smart grids using big electrical data, Cluster Comput. 22 (1) (2019) 241–270.
[128] K. Wang, et al., Robust big data analytics for electricity price forecasting in the smart grid, IEEE Transactions on Big Data 5 (1) (2019) 34–45.
[129] Y. Zhang, et al., The power big data-based energy analysis for intelligent community in smart grid, Int. J. Embed. Syst. 11 (3) (2019) 295–305.